PLOS ONE

RESEARCH ARTICLE

BiomMiner: An advanced exploratory microbiome analysis and visualization pipeline

Amirhossein Shamsaddini*, Kimia Dadkhah, Patrick M. Gillevet

Microbiome Analysis Center, George Mason University, Manassas, Virginia, United States of America

* ashamsad@gmu.edu

OPEN ACCESS

Citation: Shamsaddini A, Dadkhah K, Gillevet PM (2020) BiomMiner: An advanced exploratory microbiome analysis and visualization pipeline. PLoS ONE 15(6): e0234860. https://doi.org/10.1371/journal.pone.0234860

Editor: Lingling An, University of Arizona, UNITED STATES

Received: December 25, 2019

Accepted: June 3, 2020

Published: June 18, 2020

Data Availability Statement: Data are available at the George Mason University Microbiome Analysis Center at http://mbac.gmu.edu/mbac_wp/biomminer-readme/.

Funding: This research was supported by Dr. Gillevet's indirect funds at the Microbiome Analysis Center. The funders played no role in the study design, analysis, decision to publish, or preparation of the manuscript. There is no competing interest for any of the authors.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Current microbiome applications require substantial bioinformatics expertise to execute. As microbiome clinical diagnostics are being developed, there is a critical need to implement computational tools and applications that are user-friendly for the medical community to understand microbiome correlation with the health. To address this need, we have developed BiomMiner (pronounced as "biominer"), an automated pipeline that provides a comprehensive analysis of microbiome data. The pipeline finds taxonomic signatures of microbiome data and compiles actionable clinical report that allows clinicians and biomedical scientists to efficiently perform statistical analysis and data mining on the large microbiome datasets. BiomMiner generates web-enabled visualization of the analysis results and is specifically designed to facilitate the use of microbiome datasets in clinical applications.

Introduction

Targeted amplicon-based analysis using 16S ribosomal RNA (rRNA) gene sequences is frequently used to explore complicated bacterial communities such as the human gut microbiome [1]. This approach has been used since 2007 for clinical diagnostics [2]. Comparative metagenomics has determined that there are three major 'enterotypes' affiliated with human gut microflora, and each of these enterotypes has a signatures genus, Bacteroides in the enterotype 1, Prevotella in the enterotype 2, and Ruminococcus in the enterotype 3 [3]. Another comparative metagenomics studies revealed a different gut microflora between 'lean' and 'obese' individuals [4]. Analysis of the large and complex bacterial communities like these studies demands sophisticated bioinformatics tools to efficiently process data in order to obtain a clear understanding of the dynamics of these ecological systems. There are several applications and pipelines available to process 16S rRNA gene sequencing data. The most popular open source packages are QIIME [5] and mothur [6]. Both QIIME and mother are all self-contained pipelines which can be used to analyze 16S rRNA gene sequencing data. Due to their comprehensive features and support documentation, QIIME and mother are considered the standard applications for microbiome analysis [7, 8]. As the microbiome field is rapidly expanding, demands for extra features and new more robust algorithm is high. Additionally, there is a need to making these packages more accessible to the clinical community. For

instance, mothur and QIIME offer more than 100 individual commands and QIIME 2 has more than 15 commands and around 90 subcommands. Additionally, the installation of QIIME2 requires at least one hour using their private Bioconda [9] Channel on a high-end computer. Both QIIME and mothur documentation are detailed and include installation instructions with various tutorials that walking the reader step by step through a sequence of pipeline commands with example datasets to test their installation. While detailed documentation is helpful for research professionals, it can be overwhelming for the clinical users as they may not understand how modifications of complicated settings may alter the outcome and change the final analysis.

BiomMiner provides an advanced comprehensive data analysis workflow which covers both upstream and downstream analysis of 16S rRNA gene data sets. By eliminating the burden of command-line and step-by-step data processing, BiomMiner simplifies the processing down to single command and provides a standard HTML package of all generated downstream results with provenance logs for each step in the pipeline. This provides a simple mechanism to analyze microbiome data that is reproducible and easy to understand. This is critical to support clinical studies and the clinical diagnostics. BiomMiner offers the flexibility to choose between multiple Standard Operating Procedure (SOPs) such as mothur MiSeq SOP (https://www.mothur.org/wiki/MiSeq_SOP) for upstream data processing and provides a wide range of downstream statistical analysis with visualizations in a single HTML package. Sets of parameters are stored in JSON configuration files that can be used to reproducibly modify and re-run pipelines for evaluation and comparison using visualization within the HTML package. We provide documentation, installation instruction, example datasets, and case sample reports to facilitate rapid evaluation and adoption of the software under the MIT license at https://mbac.gmu.edu/mbac_wp/biomminer-readme/

Approach

BiomMiner uses Snakemake [10] as the primary workflow management language for scalability and reproducible execution of various wrapper scripts developed in python and R for existing software tools. BiomMiner can easily redo failed steps and resume from checkpoints without repeating computationally intensive tasks which facilitates the testing of different parameters in a workflow. The other aspect of BiomMiner is the ability to deploy on both large clusters such as Amazon Web Services (AWS) or a single desktop computer with a few cores. BiomMiner generates an HTML package as standardized output using JavaScript to catalog all generated charts, graphs, and text-based result. Most of the results are visualized using ggplot2 [11] which can generate a high-resolution image with different formats. The user can open the HTML package on Internet browsing applications such as Chrome, Firefox, or Safari. The pipeline uses a single workflow configuration file (JSON config file) that can control most of the essential steps of the workflow and the users can easily modify them based on their research goals. At each step, BiomMiner keeps track of the logs generated in a specific directory that can be used to monitor the process.

Results

BiomMiner utilizes many publicly available tools to perform the major steps of 16S rRNA analysis. Where necessary, we have written wrapper scripts to allow multiple samples to be run simultaneously and to integrate multiple tools by seamlessly converting file format. These scripts are typically written in either R or Python and are available at the BiomMiner tutorial link. BiomMiner workflow is divided into upstream and downstream pipelines. The upstream

pipeline of the BiomMiner workflow follows the Schloss lab Standard Operating Procedure (https://www.mothur.org/wiki/MiSeq_SOP) for Illumina Miseq-SOP using mothur v1.34.

The downstream part of the BiomMiner workflow executes and visualizes the most popular statistical approaches for microbiome analysis such as alpha diversity, beta diversity, machine learning and generate an HTML file including results from downstream steps at the end of execution.

BiomMiner upstream analysis modules

1. Paired read merging (assembly). If the data is generated by an Illumina instrument for example Illumina Miseq, read constructs are sequenced in both directions called "paired-end" read. BiomMiner merges pairs and creates one single read per pair generating a consensus sequence by aligning the forward and reverse reads and resolving any mismatches found in the alignment. This is accomplished by the "make.contigs" command in the mothur package.
2. Reducing sequencing and PCR errors. Raw reads that are generated by a next-generation sequencing machines such as 454 or Illumina have predicted error probabilities for each base indicated by quality (Q) scores. In many applications it is important to filter low quality reads to reduce the number of errors, especially in 16S rRNA gene sequencing experiments. The mothur "screen.seqs" command is used to filter out low-quality reads.
3. Chimera detection and removal. Chimeric sequences are an artifact formed from two or more different sequences joined together during PCR amplification. Chimeras are rare with shotgun sequencing but are common in amplicon sequencing when closely related sequences are amplified. The "chimera.vsearch" command is used to detect and discard chimeric reads.
4. Dereplication. The pipeline then discard duplicate sequences by running "uniq.seqs" command in the mothur package which compares every base in a sequence read and they must be identical over the full length of both sequences to be consider as duplicates.
5. Cluster the sequences into OTUs. We then use clustering algorithm (mothur opticlust) to create groups of closely related reads based on the similarity threshold (97% similarity) called operational taxonomic unit (OTU).
6. Assign taxonomic annotation to each OTU. We use RDP [12] v.16 as a reference in the command "classify.otu" to assign a consensus taxonomy for each OTU.
7. Create an OTU abundance table. OTU abundance tables are often stored as tabbed text files in which OTUs are rows and samples are columns. The abundance of an OTU is the number of reads derived from all biological sequences that are $>=$ 97% identical to the OTU sequence. One entry in the table is usually a number of reads, also called a "count" or can be converted to relative abundance in the range 0.0 to 1.0.

BiomMiner downstream analysis modules

BiomMiner starts processing the OTU abundance table by generating a comprehensive HTML report including several most popular statistical approaches for microbiome analysis. These include an overview, alpha diversity, differential abundance analysis, beta diversity, and machine learning as shown in Fig 1.

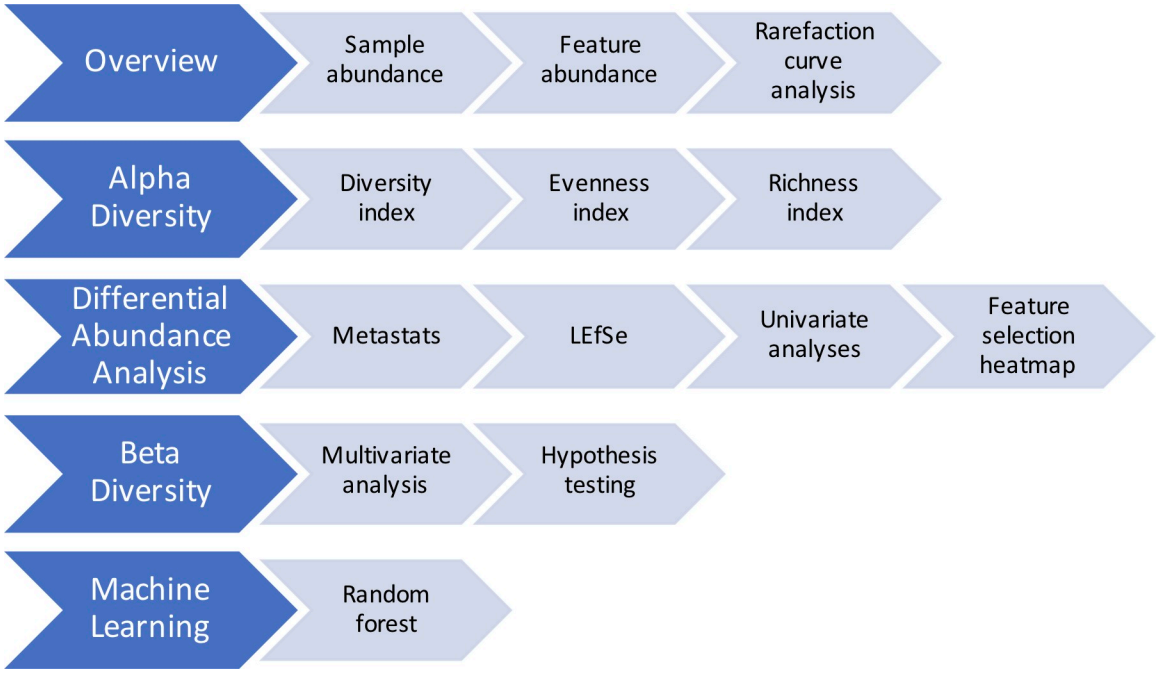

Fig 1. BiomMiner downstream analysis modules. There are five downstream analysis modules for BiomMiner that all start with the OTU abundance table as the input.

https://doi.org/10.1371/journal.pone.0234860.g001

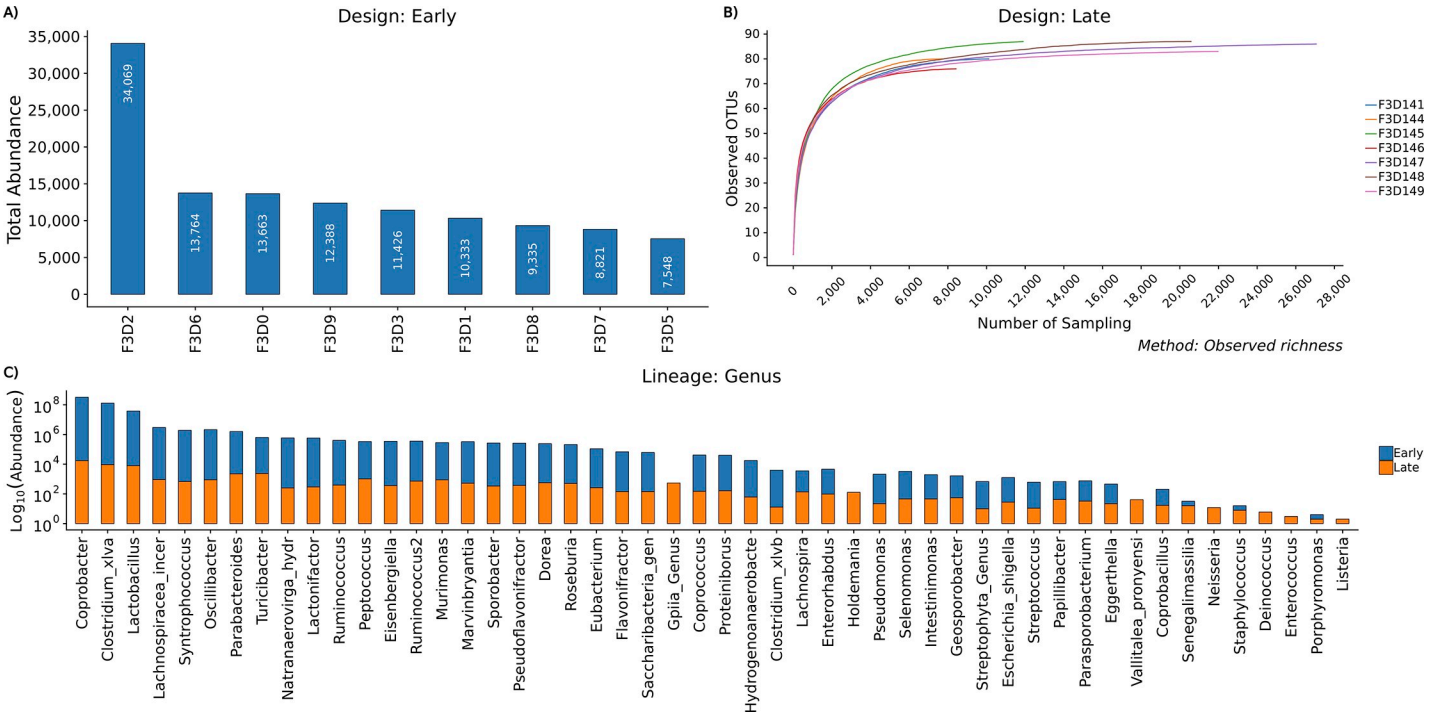

Fig 2. BiomMiner overview. Selected benchmark is used here to calculate and visualize the significant changes in the murine gut microbiome community in two state communities of the study, Early (10 days following weaning) and Late (15 days following weaning) [13]. This module produces total sample abundance bar plot, rarefaction curve, and groupwise feature abundance bar plot. Images here are from the study that is used as benchmark. (A) Total sample abundance bar chart for each community. Only the early group is shown here. The X-axis represents the sample name of the condition, and the Y-axis represents the total abundance of each sample. (B) Rarefaction curve plot. Only Late group is shown here. The rarefaction curve of the Late group reached an asymptote, which indicated that the sequencing depth was sufficient to represent the majority of species richness (observed richness). The X-axis represents number of samplings without replacement and the Y-axis represents the number of unique observed OTUs. (C) Total Feature Abundance bar chart. Log scaled comparison of the most abundant phylotypes between Early and Late group at the genus level. The X-axis represents Genus-level taxon and the Y-axis represents the abundance of each genus-level taxon on a log10 scale.

https://doi.org/10.1371/journal.pone.0234860.g002

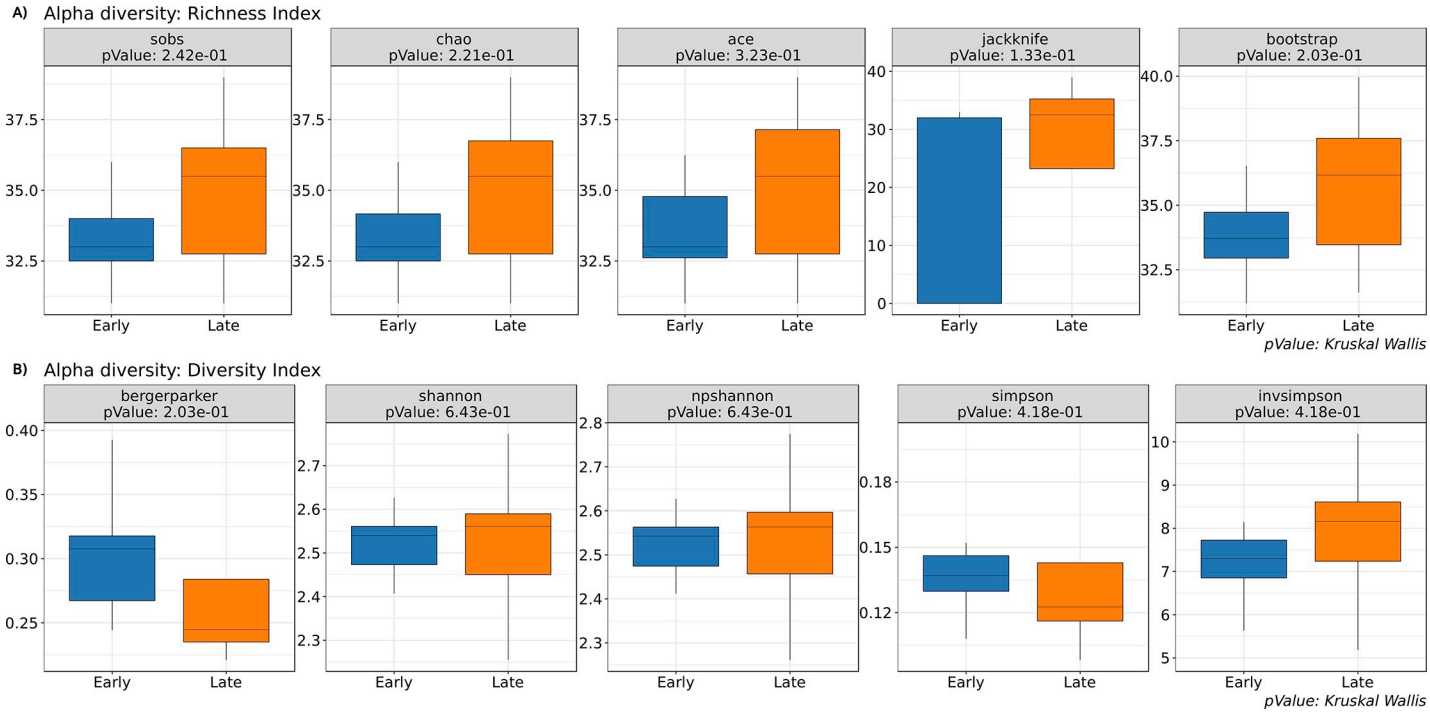

Fig 3. Alpha diversity. The selected benchmark is used here to calculate and visualize the significant changes in the murine gut microbiome community in the two states of the study, Early (10 days following weaning) and Late (15 days following weaning) [13]. Box and whiskers plots illustrate the median, quartiles, maximum and minimum of the alpha diversity value based on specific metrics. *pvalue* indicates significant difference between groups using Kruskal Wallis test. (A) Richness index boxplot. Richness index including sobs(Observed richness), Chao1 [17], ACE [18], Jackknife and bootstrap [19] were used to identify community richness differences between two groups. The X-axis represents biological condition and the Y-axis represents distribution of calculated richness index.(B) Diversity index boxplot. Diversity index including Berger-parker [16], Shannon, non-parametric Shannon [14], Simpson, and inverse-Simpson [15] were used to identify community diversity for diversity between subgroups of Early and Late category. Based on the box plots, there were no differences in community diversity between study subjects. The X-axis represents biological condition and the Y-axis represents the distribution of calculated diversity index.

https://doi.org/10.1371/journal.pone.0234860.g003

A public benchmark from Schloss et al. 2012 [13] is selected here to represent the analysis pipeline and visualization features of BiomMiner. The selected benchmark is used to understand the effect of normal variation in the gut microbiome on host health. The study has been developed to determine whether there were significant changes in the murine gut microbiome community during the first year of life in Early (10 days following weaning) and Late (15 days following weaning).

Overview

The main aim of the overview module is to provide a summary of the generated OTU abundance table like Groupwise sample abundance, feature abundance total count, and Rarefaction curve analysis. Groupwise sample abundance displays the total abundance of each sample for each Biological Condition as a Bar chart. Feature Abundance displays the total count of each OTU per each community which describes the distribution of OTU abundance per community. Rarefaction curve analysis, the estimate of sequencing depth and richness for each sample, is a very popular metric in microbiome analysis. BiomMiner uses mothur v.1.34 to perform rarefaction analysis. The goal of rarefaction is to determine whether sufficient observations have been made to get a reasonable estimate of a quantity that has been measured by sampling. The most commonly considered quantity is OTU richness (the number of different OTU in a group) (Fig 2).

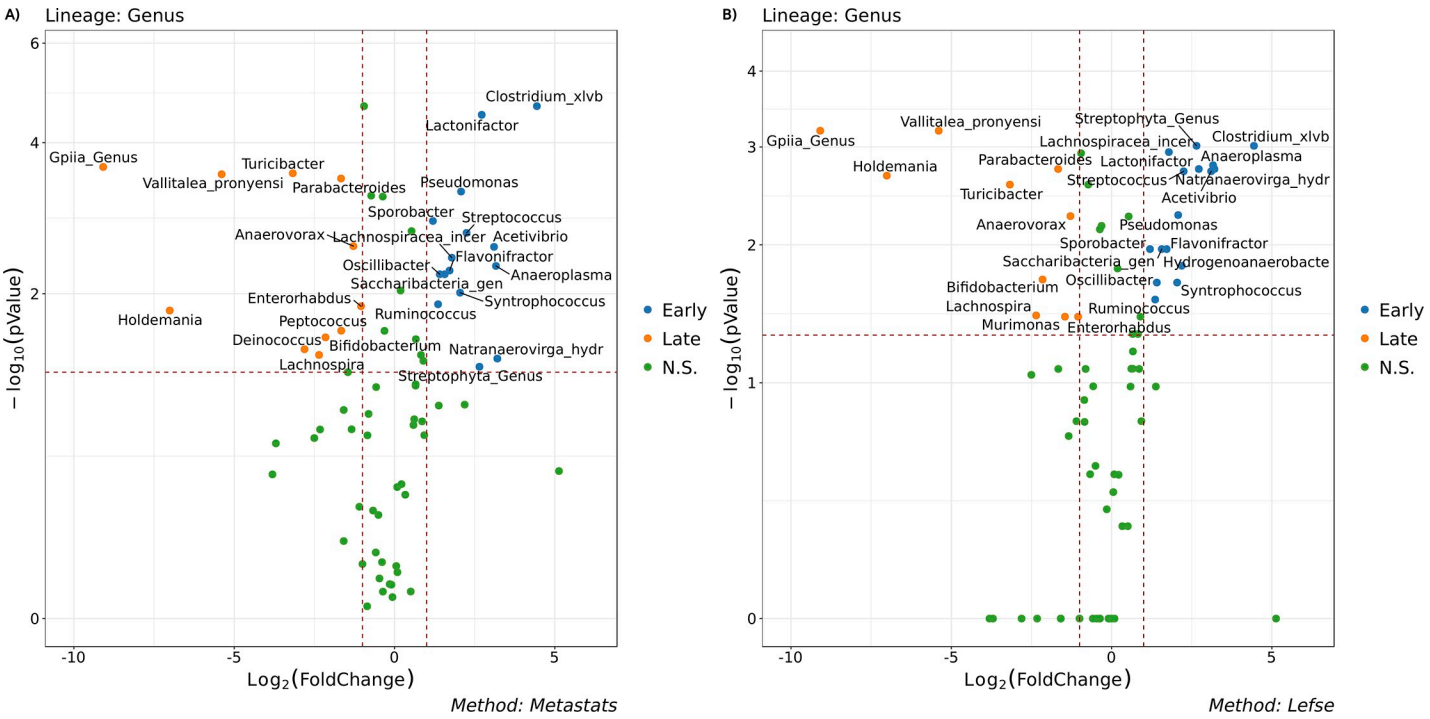

Fig 4. Differential abundance analysis: Volcano plot. The selected benchmark is used here to calculate and visualize the significant changes in the murine gut microbiome community in the two state communities of the study, early (10 days following weaning) and late (15 days following weaning) [13].Volcano plot showing OTU fold changes on X-axis and the negative logarithm (base 10) of the Bonferroni-adjusted *pvalue* on Y-axis. Dashed vertical and horizontal lines reflect the filtering criteria (fold change = ±1.0 and Bonferroni-adjusted *pvalue* > -log (0.05). Blue or Orange dots represent Genus entities that are significant based on Specific test (LEfSe or Metastats) at each group. The Green dots (N.S.) represent the Genus features either common between groups or classified as insignificant by the test (LEfSe or Metastats). In both A and B plots, the X-axis represents the abundance fold change on log2 scale, and the Y-axis represents the negative log10 of the calculated *pvalue*. N. S. means Non-significant. (A) Metastats Volcano plot suggest the differential features in metagenomic across two studies. (B) LEfSe Volcano plot could be interpreted as consistent difference in relative abundance of the analyzed fecal bacteria communities across the two groups.

https://doi.org/10.1371/journal.pone.0234860.g004

Alpha diversity

Alpha diversity is the diversity within an individual sample. There are several alpha diversity indices available in BiomMiner to investigate diversity, richness and evenness such as Shannon [14], Simpson [15], Berger-Parker [16], and chao1 [17]. We are using mothur v.1.34 to calculate alpha diversity estimate. The richness estimate indicates the number of OTU found in a given sample regardless of how common or rare they are. The Evenness estimate indicates how evenly the richness (OTU count) is distributed. The Diversity estimate is a measurement of richness combined with evenness meaning it takes into account not only how many OTU is present but also how evenly distributed the numbers of each OTU are. After calculating Alpha diversity value of each population, BiomMiner then uses the calculated alpha diversity estimates in a statistical test to check whether the diversity, richness, and evenness between two conditions are significantly different by calculating the Bonferroni corrected *p*value of a Kruskal-Wallis test (Fig 3).

Differential abundance analysis

BiomMiner also performs statistical methodology designed to identify differentially abundant features in metagenomic and 16S rRNA sequence datasets. We utilize well-established methods such as Metastats [20], LEfSe [21], and Kruskal–Wallis test available in mothur v.1.34. Metastats perform Fisher's exact test and calculate *pv*alue to provide a list of interesting features that are different between two groups. LEfSe (Linear discriminant analysis Effect Size)

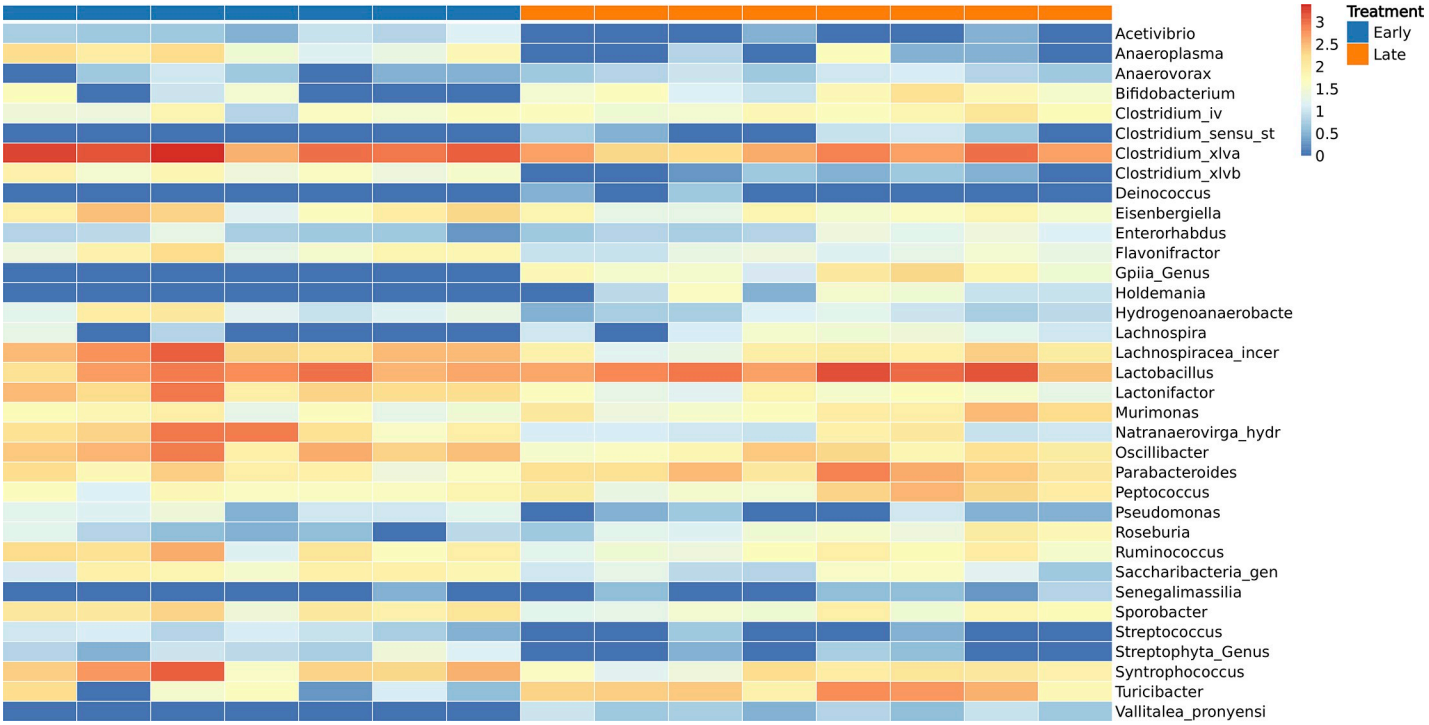

Fig 5. Differential abundance analysis: Heatmap. The selected benchmark is used here to calculate and visualize the significant changes in the murine gut microbiome community in two state communities of the study, early (10 days following weaning) and late (15 days following weaning) [13]. Heatmap showing the abundance variation of top 35 common bacterial taxa at the genus level which were significant OTUs (pvalue < 0.05) based on LEfSe, Metastats, and Kruskal-Wallis test. The rows represent the bacterial taxa and columns are the samples.

https://doi.org/10.1371/journal.pone.0234860.g005

selects features (OTU) most likely to explain differences between communities by coupling a Kruskal–Wallis and a Wilcoxon rank-sum test) for statistical significance with a Linear Discriminant Analysis (LDA) to define the effect relevance.

Kruskal-Wallis (one-way ANOVA on ranks) is a non-parametric method for testing whether features originate from the same distribution between two communities. To quickly identify changes in large data, we used the "volcano plot" to present the result of each test. It is scatter plot which plots magnitude of the change(foldchange) of an OTU versus *pv*alue of the OTU from a statistical test on the X and Y-axis respectively, enabling quick visual identification of most the important features. To be considered as a significant feature; the foldchange value should be greater than 1 or less than -1, and the negative logarithm (base 10) of the *pv*alue should be above 1.13(-log (0.05)). We color each point based on their foldchange and *pv*alue so the user can easily pinpoint the biological and statistical significance of OTUs (Fig 4).

Since different statistical models sometimes produce *pv*alues that can be vastly different from each other, it is advisable to compare results from multiple methods and to visualize the features to gain more confidence. BiomMiner selects features by combining common significant features from each statistical test and then displays up to top 50 of these distinctive features in a heatmap plot. This implementation allows users to clearly pinpoint features of interest while minimizing the chance of missing important ones (Fig 5).

Beta diversity: Ordination analysis

Ordination measurements are used to compare the similarity/ dissimilarity of the microbial communities. Microbiome studies are typically sparse with high-dimensionality, so it is hard

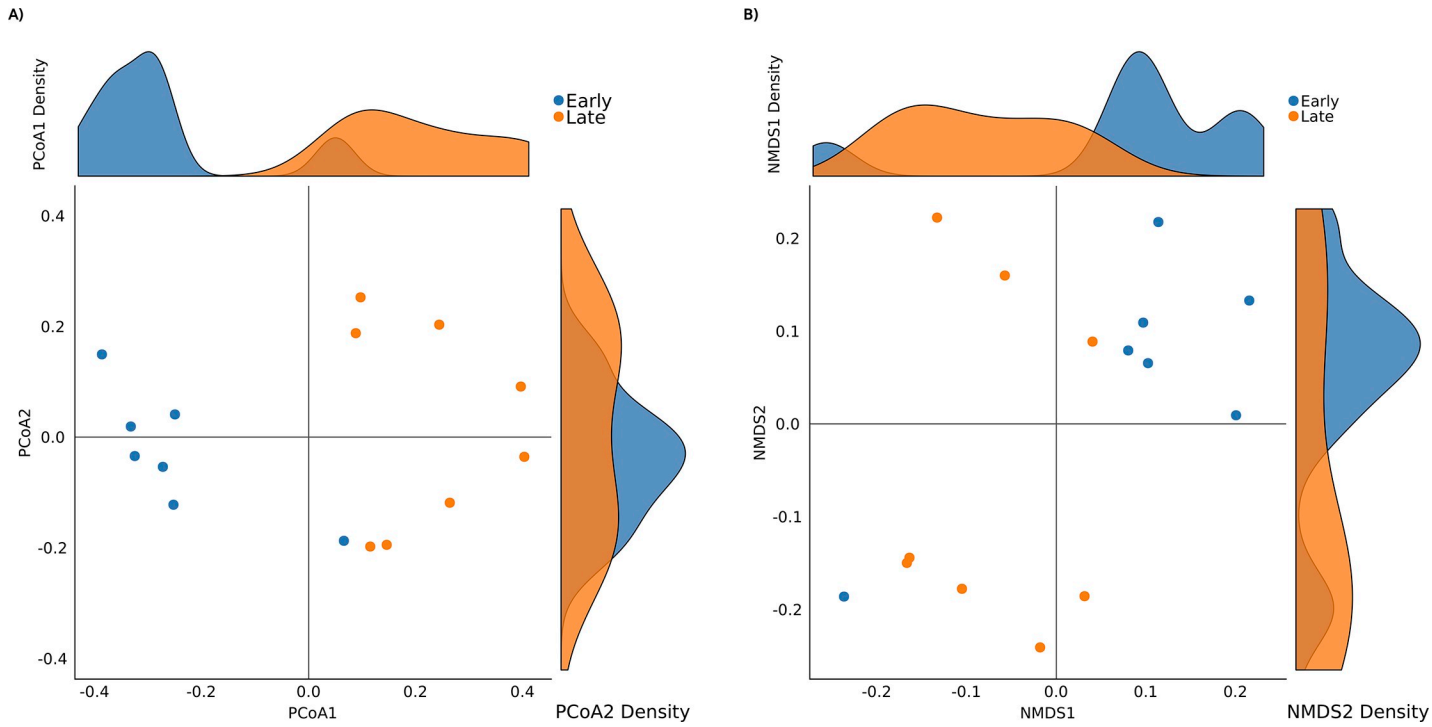

Fig 6. Beta diversity: Ordination analysis. The selected benchmark is used here to calculate and visualize the significant changes in the murine gut microbiome community in two state communities of the study, early (10 days following weaning) and late (15 days following weaning) [13]. Ordination analysis of microbial communities of the two groups was calculated using Bray-Curtis distance matrix and visualized using principal coordinate analysis (PCoA) (Plot A) and non-metric multidimensional scaling (NMDS) (Plot B). Points represent samples. Samples that are more similar to one another are ordinated closer together. The density plot on axis can be used to identify the similarity of distribution. The X-axis represent the first axis of the ordination while displaying the density of the sample's ordination on first axis, and the Y-axis represent the second axis of the ordination while displaying the density of the sample's ordination on second axis.

https://doi.org/10.1371/journal.pone.0234860.g006

to assess the direct correlation of microbiome composition with potential biological factors using OTUs abundances. Thus, ordination analyses is generally used to select a distance measurement method between groups and then conduct an analysis of the estimated distances [22]. BiomMiner utilize mothur v.1.34 for Beta diversity analysis. BiomMiner's ordination module uses popular distance measure algorithms in microbiome studies like BrayCurtis [23], Jaccard [24], and weighted/ unweighted Unifrac [25] for performing ordination analysis and hypothesis testing to evaluate the dissimilarity of microbial community in each distance matrix.

Ordination plots the distances between the communities into a Euclidean space and are then visualized via principal-coordinate analysis (PCoA) or non-metric multidimensional scaling (NMDS). Given a matrix of distances between samples, a PCoA visualizes these in a 2-dimensional Euclidian space represents their pair-wise distance in the original matrix. Non-metric multidimensional scaling (NMDS) is an indirect gradient analysis approach which produces an ordination based on a distance or dissimilarity matrix [26]. NMDS attempts to represent, as closely as possible, the pairwise dissimilarity between objects in a low-dimensional space (Fig 6).

Beta diversity: Statistical hypothesis test

The Beta diversity's statistical null hypothesis in microbiome studies is developed as "there is no difference of microbiome composition in experimental groups (e.g., healthy vs. patient)" or "there is no differences in distribution or structure of population of microbiome between cohorts". BiomMiner uses most common approaches of microbiome hypothesis testing

A)

Test	PERMANOVA(ADONIS)
Hypothesis	Distributions of population is similar
Statistics	pseudo-Fisher's Test
Distance	Jaccard
Sample Size	15
Group Size	2
Groups	Early-Late
Permutations	1000
test statistic	6.433
p-value	<0.001***

B)

Test	ANOSIM
Hypothesis	the similarity between groups is greater than or equal to the similarity within the groups
Statistics	mean Rank
Distance	Jaccard
Sample Size	15
Group Size	2
Groups	Early-Late
Permutations	1000
Early-Late	pValue: <0.001*

C)

Test	AMOVA
Hypothesis	Distribution of Variance is similar
Statistics	Fisher's Test
Distance	Jaccard
Sample Size	15
Group Size	2
Groups	Early-Late
Permutations	1000
Early-Late	p-value: <0.001*

D)

Test	LIBSHUFF
Hypothesis	the population structure is similar
Statistics	Cramer-von Mises statistic
Distance	Jaccard
Sample Size	15
Group Size	2
Groups	Early-Late
Permutations	1000
Early-Late	Pvalue: 0.022
Late-Early	Pvalue: 0.022

Fig 7. Beta diversity: Statistical hypothesis test. The selected benchmark is used here to calculate and visualize the significant changes in the murine gut microbiome community in two state communities of the study, early (10 days following weaning) and late (15 days following weaning) [13]. We used Jaccard distances as input for statistical hypothesis tests, comparing microbial community composition between two groups.(A) Permutational multivariate analysis of variance (PERMANOVA) compare microbial community and test the null hypothesis that distribution of microbial population is similar.(B) The Analysis of similarity (ANOSIM) is a nonparametric analog of traditional analysis of variance (ANOVA) and compares the mean of ranked dissimilarities between groups to the mean of ranked dissimilarities within groups.(C) Analysis of molecular variance (AMOVA) tests the variance of distribution between two groups.(D) LIBSHUFF describes whether two or more communities have the same structure using the Cramer-von Mises test statistic.

https://doi.org/10.1371/journal.pone.0234860.g007

methods like AMOVA [27], HOMOVA [28], ANOSIM [29], LIBSHUFF [30], and PERMANOVA [31], and then displays the result of each test including the details of the test in separate table (Fig 7). The Analysis of similarity (ANOSIM) is rank-based or nonparametric version of analysis of variance (ANOVA) uses dissimilarity matrixes to provides a single *p*value indicating if community profiles (OTUs) similarity are significantly different between groups. PERMANOVA (Adonis) is a multivariate technique analogous to MANOVA and describes whether the variation in community OTU's composition is different between groups. AMOVA (Analysis of Molecular Variance) can be used to measure the apportionment of OTU variance between pairs of populations [32]. LIBSHUFF describes whether two or more communities have the same OTU structure, different, or subsets of one another using the Cramer-von Mises test statistic. Homogeneity of molecular variance (HOMOVA) determines whether the diversity of features (OTUs) in each community is significantly different.

Machine learning

In addition to the standard statistical approaches mentioned above, BiomMiner also supports a number of machine learning approaches for supervised learning and feature selection, such as random forest (RF) and support vector machine (SVM). In many recent reports on the classification of microbiome data, it has been shown that machine learning and data mining have performed well [33, 34]. BiomMiner uses the Caret [35] R package to calculate Random forest (RF) and SVM and uses the "predict" R package modeling algorithm on the test set. When

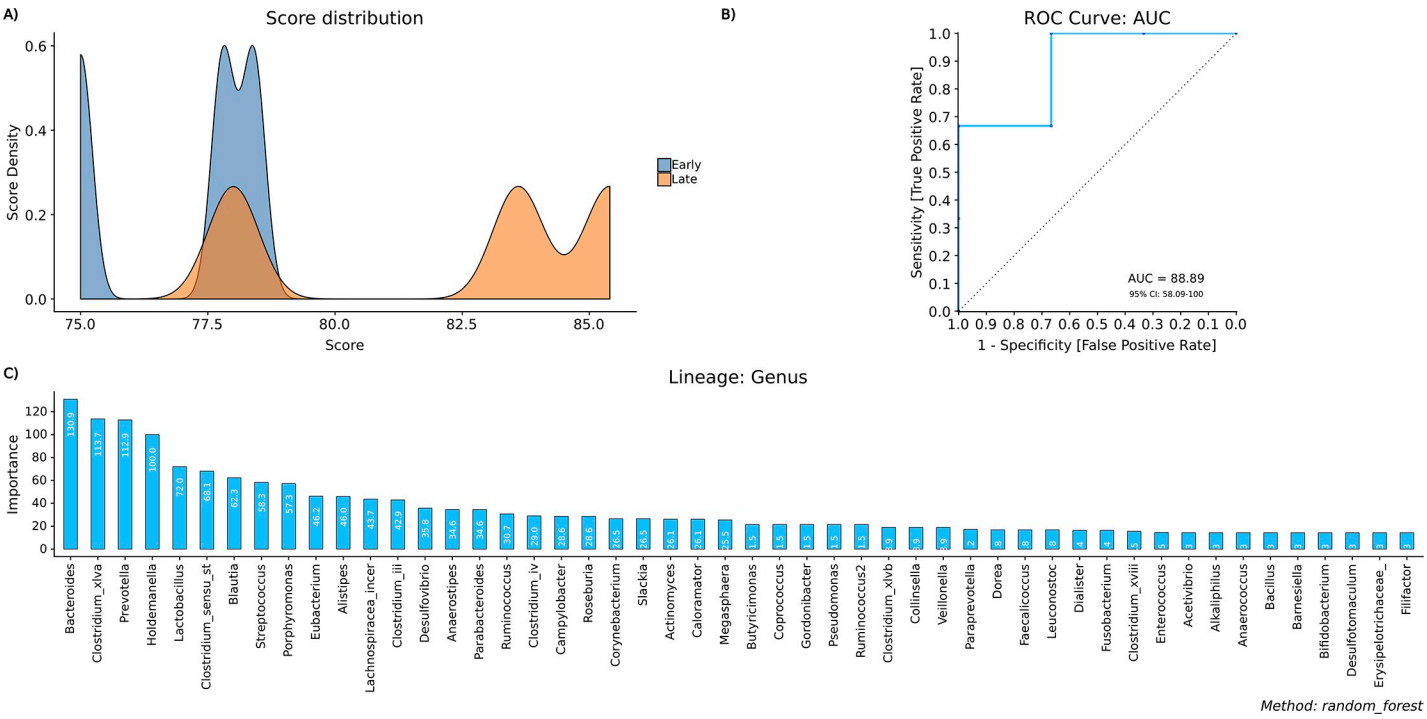

Fig 8. Machine learning module. The selected benchmark is used here to calculate and visualize the significant changes in the murine gut microbiome community in two state communities of the study, early (10 days following weaning) and late (15 days following weaning) [13]. (A) Score distribution gives us visual information on skewness, distribution and our model's facility to distinguish each class. Here we can see how the model has distributed both our categories, (the more separate, the better). The X-axis represent the distributions of calculated prediction score and the Y-axis represent the sample prediction score density. (B) Receiver operating characteristic (ROC) curves for Random Forest classifier. The ROC curve will give us an idea of how our model is performing with our test set. The ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) where TPR is on Y-axis and FPR is on the X-axis. if the AUC is close to 50% then the model is as good as a random selector. On the other hand, if the AUC is near 100% then you have a "perfect model". (C) Important feature. It shows the importance of each feature by calculating the increase in the model's prediction error after permuting the feature. A feature is "important" if shuffling its values increases the model error as the model relied on the feature for prediction accuracy. A feature is "unimportant" if shuffling its values leaves the model error unchanged as the feature did not contribute to the prediction accuracy.

https://doi.org/10.1371/journal.pone.0234860.g008

running machine learning in BiomMiner we used 70% of OTU abundance table as the training set to train the model and to evaluate the performance of the generated model and the remaining 30% of the OTU abundance table as a test set. In order to assess the performance of the machine learning model, we plot the Receiver operating characteristic (ROC) curve, the predicted score distributions density plot, and the important feature bar plot. The Receiver operating characteristic (ROC) curve indicates how well our model is performing with our test set. It tells how much of the model is capable of distinguishing between communities. The ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) where TPR is on Y-axis and FPR is on the X-axis. The predicted score distributions density plot gives us visual information on skewness, distribution, and our model's accuracy to distinguish each class. Important feature bar plot shows the impact of each feature (OTU) on the accuracy of the model. A feature (OTU) is "important" if shuffling its values increases the model error indicating that the model relied on the feature for the prediction. Clearly, for unimportant OTU, the permutation of its value will have little to no effect on the model accuracy (Fig 8).

Discussion

Several excellent web-based or desktop applications have been developed over the past decade to support microbiome data analysis. Most of these tools have been developed primarily for

raw sequence processing, clustering, and annotation, with limited or yet in development support for advanced statistical analysis and visual exploration. Other applications only focused on the downstream portion of the analysis and let the user upload their processed data which suffer from several issues like format incompatibility, unsupported annotation (which may lead to garbage in, garbage out patterns). BiomMiner complements these applications by providing complete upstream analysis and comprehensive support for statistical, visual, and meta-analysis on the downstream side of the experiment. While developing BiomMiner, we aimed to create a sophisticated yet easy to understand platform for microbiome data analysis. Users can easily download the analysis result at high-resolution images that generated on BiomMiner for using in their publications or they can import text-based results from BiomMiner into other software for further analyses. The future advancement of BiomMiner will focus on integrating new downstream analysis such as functional genomics.

Supporting information

S1 File. Summary of comparison between BiomMiner and currently available software. This PDF file contains a table summarizing a comparison of supported capabilities between BiomMiner, phyloseq [1], QIIME and mothur. A "1" or "0" indicates that the capability is supported or not supported, respectively. "T" means the result is available as text format file and "G" indicate the generated file is graphic based result and "T/G" mean the result is available in text and graphic based format. This is not a comprehensive summary of the capabilities of each package, but rather the capabilities of relevance to this article.
(PDF)

S1 Data.
(DOCX)

Author Contributions

Conceptualization: Amirhossein Shamsaddini, Kimia Dadkhah, Patrick M. Gillevet.

Data curation: Amirhossein Shamsaddini.

Formal analysis: Amirhossein Shamsaddini.

Investigation: Amirhossein Shamsaddini.

Methodology: Amirhossein Shamsaddini, Kimia Dadkhah.

Project administration: Patrick M. Gillevet.

Software: Amirhossein Shamsaddini.

Supervision: Patrick M. Gillevet.

Validation: Patrick M. Gillevet.

Visualization: Amirhossein Shamsaddini.

Writing – original draft: Amirhossein Shamsaddini.

Writing – review & editing: Kimia Dadkhah, Patrick M. Gillevet.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007; 449(7164):804–10. https://doi.org/10.1038/nature06244 PMID: 17943116

2. Komanduri S, Gillevet PM, Sikaroodi M, Mutlu E, Keshavarzian A. Dysbiosis in pouchitis: evidence of unique microfloral patterns in pouch inflammation. Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association. 2007; 5(3):352–60. Epub 2007/03/21. https://doi.org/10.1016/j.cgh.2007.01.001 PMID: 17368235.

3. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011; 473(7346):174–80. https://doi.org/10.1038/nature09944 PMID: 21508958

4. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006; 444(7122):1027–31. https://doi.org/10.1038/nature05414 PMID: 17183312

5. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. 2010; 7(5):335–6. https://doi.org/10.1038/nmeth.f.303 PMID: 20383131

6. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology. 2009; 75(23):7537–41. https://doi.org/10.1128/AEM.01541-09 PMID: 19801464

7. Nilakanta H, Drews KL, Firrell S, Foulkes MA, Jablonski KA. A review of software for analyzing molecular sequences. BMC Research Notes. 2014; 7(1):830. https://doi.org/10.1186/1756-0500-7-830 PMID: 25421430

8. Plummer E, Twin J. A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. Journal of Proteomics & Bioinformatics. 2015; 8 (12). https://doi.org/10.4172/jpb.1000381

9. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods. 2018; 15(7):475–6. https://doi.org/10.1038/s41592-018-0046-7 PMID: 29967506

10. Koster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012; 28(19):2520–2. https://doi.org/10.1093/bioinformatics/bts480 PMID: 22908215

11. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer; 2009. 1–212 p.

12. Larsen N, Olsen GJ, Maidak BL, McCaughey MJ, Overbeek R, Macke TJ, et al. The ribosomal database project. Nucleic Acids Research. 1993; 21(13):3021–3. https://doi.org/10.1093/nar/21.13.3021 PMID: 8332524

13. Schloss PD, Schubert AM, Zackular JP, Iverson KD, Young VB, Petrosino JF. Stabilization of the murine gut microbiome following weaning. Gut Microbes. 2012; 3(4):383–93. https://doi.org/10.4161/gmic.21008 PMID: 22688727

14. Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal. 1948; 27 (3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x PMID: 30854411

15. Simpson EH. Measurement of Diversity. Nature. 1949; 163(4148):688-. https://doi.org/10.1038/163688a0

16. Berger WH, Parker FL. Diversity of Planktonic Foraminifera in Deep-Sea Sediments. Science. 1970; 168(3937):1345–7. https://doi.org/10.1126/science.168.3937.1345 PMID: 17731043

17. Chao A, Chazdon R, Colwell R, Shen T-J. Chao A, Chazdon RL, Colwell RK, Shen T-J. A new statistical approach for assessing compositional similarity based on incidence and abundance data. Ecol Lett 8: 148–159. Ecology Letters. 2005; 8:148–59. https://doi.org/10.1111/j.1461-0248.2004.00707.x

18. Chao A, Lee S-M. Estimating the Number of Classes via Sample Coverage. 1992; 87(417):210–7. https://doi.org/10.1080/01621459.1992.10475194

19. Efron B. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika. 1981; 68(3):589–99. https://doi.org/10.1093/biomet/68.3.589

20. White JR, Nagarajan N, Pop M. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS Computational Biology. 2009; 5(4):e1000352. https://doi.org/10.1371/journal.pcbi.1000352 PMID: 19360128

21. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biology. 2011; 12(6):R60. https://doi.org/10.1186/gb-2011-12-6-r60 PMID: 21702898

22. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Genes & Diseases. 2017; 4 (3):138–48. https://doi.org/10.1016/j.gendis.2017.06.001 PMID: 30197908

23. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs. 1957; 27(4):325–49. https://doi.org/10.2307/1942268

24. Jaccard P. Etude de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Societe Vaudoise des Sciences Naturelles. 1901; 37:547–79. https://doi.org/10.5169/seals-266450

25. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and Qualitative Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. Applied and Environmental Microbiology. 2007; 73(5):1576–85. https://doi.org/10.1128/AEM.01996-06 PMID: 17220268

26. Buttigieg PL, Ramette A. A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. FEMS Microbiology Ecology. 2014; 90(3):543–50. https://doi.org/10.1111/1574-6941.12437 PMID: 25314312

27. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 1992; 131 (2):479–91. PMID: 1644282

28. Stewart CN, Excoffier L. Assessing population genetic structure and variability with RAPD data: Application to Vaccinium macrocarpon (American Cranberry). Journal of Evolutionary Biology. 1996; 9 (2):153–71. https://doi.org/10.1046/j.1420-9101.1996.9020153.x

29. Clarke KR. Non-parametric multivariate analyses of changes in community structure. Austral Ecology. 1993; 18(1):117–43. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x

30. Singleton DR, Furlong MA, Rathbun SL, Whitman WB. Quantitative Comparisons of 16S rRNA Gene Sequence Libraries from Environmental Samples. Applied and Environmental Microbiology. 2001; 67 (9):4374–6. https://doi.org/10.1128/aem.67.9.4374-4376.2001 PMID: 11526051

31. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001; 26(1):32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

32. Meirmans PG, Liu S. Analysis of Molecular Variance (AMOVA) for Autopolyploids. Frontiers in Ecology and Evolution. 2018; 6(66). https://doi.org/10.3389/fevo.2018.00066

33. Dadkhah E, Sikaroodi M, Korman L, Hardi R, Baybick J, Hanzel D, et al. Gut microbiome identifies risk for colorectal polyps. BMJ Open Gastroenterology. 2019; 6(1):e000297. https://doi.org/10.1136/bmjgast-2019-000297 PMID: 31275588

34. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiology Reviews. 2011; 35(2):343–59. https://doi.org/10.1111/j.1574-6976.2010.00251.x PMID: 21039646

35. Kuhn M. Building Predictive Models in R Using the caret Package. 2008. 2008; 28(5):26. Epub 2008-09-23. https://doi.org/10.18637/jss.v028.i05